基于大数据的交通关联事件研究

1 引言

当前交通规则将普通道路交通的实体分为三类:机动车、非机动车和行人，并对三者给出了相应的规则。但如何从频发的事件中获得相应的知识，用于指导三者的行进方式从而规避交通事件发生是一个值得研究的问题。这是一个数据挖掘问题[1]，首先要有充分交通事件的数据信息，其次要有对信息进行处理的数学理论。随着大数据产业的发展[2][3]，从中可获得大样本的交通事件信息;处理这类问题的数学方法之一是关联理论[4]。结合大数据和关联算法提出一种基于SQL例程[5][6]的交通关联事件信息的分析方法，经过对大量这类事件的数据分析获取规律性的知识，给出三实体行进时应特别注意的情况、需改变的习惯及需对交规进行修改的建议。

2 交通关联问题

由三类实体之间引发的常见关联交通问题如下：

1) 预测不足[7]:驾车者对右侧的行进体预测不足;
2) 车辆失控[8-9]:由于驾车者操作失误所至;
3) 驾车盲区[10]:与行进体同向行进右拐时，因预留空间不足所至;
4) 夜间行驶[11]:由于照明程度所至;
5) 驾车报复[12]:故意行为;
6) 行进跟踪[13]:由于分散精力所至;
7) 突生障碍:抛物、突然停车且开右侧车门等;
8) 抢夺事件:利用各类实体之间的速度差异，进行动态抢夺。

3 数据源

2015年国发《促进大数据发展行动纲要》系统部署大数据发展工作。提出2017年形成跨部门数据资源共享共用格局，当前各部门的数据平台可以认为是大数据的子集[14][15]。交通安全的大数据平台的作用在于从大量数据中合理收集、管理、处理有价值的数据，可获得对安全有助的信息。

根据大数据的4V特征[2]，可以很好的理解和应用大数据平台：

数据量大（Volume）和有价值（Value）：要从数据量大的平台上获得有价值的数据针对性要强，即要历史地、广泛地收集上述 8 类问题所至事件的相关数据。

高速获取数据（Velocity）和多种类格式（Variety）：可高速获取的数据但其格式类型众多，所以要将收集的数据经格式转换为交通事件的数据源，即将诸多非结构化数据根据算法的要求转换为相应的结构化数据以便分析。对交通事件的数据,需将现场记录、监控、责任认证等非结构化数据转换为由符号替代的有格式数据[14]，从而形成关系表，即关系数据库的基本操作对象。

以当前交通部门的数据平台为主[15]，部门官网的网上的事件发布、分析、广网评议等方式获取的交通事件信息为辅共同形成了支持挖掘的数据源。

4 算法

三类实体之间出现的事件是一个关联问题。目前，支持数据关联挖掘的数学理论是关联规则挖掘算法(Apriori algorithm)[4] [16]，该算法是在一个数据项集中找出项与项之间的关系,其核心思想是依据支持度和置信度两个定义通过频繁项集生成和检测两个阶段逐级实现挖掘。

定义 1(支持度) 支持度是测试集包含 X 的数量与总样本数量 D 之比:

$supp(X)=occur(X)/count(D)=P(X)$。如 $P(A\cap B)$表示既有 A 又有 B 的概率。

定义 2(置信度) 置信度是指测试集中包含 X 和 Y 的数与包含 X 的数量之比:

conf(X→Y)=supp(X∪Y)/supp(X)=P(Y|X)。如 P(B|A)=P(AB)/P(A), 在 A 发生的事件中同时发生 B 的概率。

Apriori 算法通过产生频繁项集和依据该集产生规则的逐层搜索的迭代方法：首先找出满足支持度或置信度的集合频繁集(frequent itemset)；其次从频繁集中找得既满足最小支持度且满足最小置信度的强规则(strong rules)。其主要瓶颈在于递增寻找候选项集，另外算法还有一个较耗时较大问题就是涉及大量 I/O 操作的数据库扫描[16]。所以优化搜索方式、与数据库的交互是改进该算法的关键。为此，提出一种基于关系表处理的改进算法:

1)将原始数据集(Itemset)中的每类项定义为表的属性名、每个 Tid 标示为一个记录。通过属性值{1,0}给出原集中项与项的关系，进而将该集等价地转换为关系表。如经典的数据项集，见表 1，经等价变换为关系表，见表 2；

表 1 数据集

Tab.1 Itemset

Tid	Items
1	A、C、D
2	B、C、E
3	A、B、C、E
4	B、E

表 2 等价关系表

tab.2 Equivalence relation table

N	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1

2)扩展这类关系表的属性值为多种取值，使应用更广范；

3)对表中的每个属性测试给出其可关联度(见定义 3)，以关联度的强度择优属性进入测试区；

4)对进入测试区的属性采用关系数据库所支持的例程计算属性间的支持度、置信度。

定义 3(可关联度) 设属性 X 中的属性值为{x1,x2,…, xj},可关联度是对属性中各属性值的一个测试计算，得出该属性中数量最多的属性值，将其的数量与该属性中全部的属性值类别数量之比值：

rltd(X)=Maxcount(xk)/Ccount(k)=Maxcount(xk)/j (k=1,2,…,j)。其中 Maxcount(xk)求得属性 X 中属性值数量最大者的数量;Ccount(k) 求得属性 X 中属性值类别的数量，其值为 j。rltd(I)值越大说明属性 X 相对于其它属性的可关联程度越高。

例如对表 2 中各属性的可关联度测试如下：

rltd(A)=Maxcount(A=1)/Ccount({0,1})= count(1)/2=2/2=1;

rltd(B)=Maxcount(B=1})/Ccount({0,1})= count(1)/2=3/2=1.7;

类似地，rltd(C)=1.7；rltd(D)=0.5；rltd(E)=1.7。得 B、C、E 是优选属性，而 A 次之，D 最差。这样只需对 B、C、E 进行可信度测试,有：

Sel @s=count(B)/count(*) from Tab2 where (B=1)and(C=1)and(E=1)

式中可信度@s=2/4=0.5=50%,这与经典算法所得的结论相同。

定义 1、2 给出了项与项之间的关联测试方法；而定义 3 给出单个属性的可关联性，是从属性的角度衡量其具有可关联程度,算法是用关联度约束数据的选择范围且对数据表进行选择操作获得结论。该算法与传统 Apriori 算法原理上的区别是以可关联度缩小查找范围而非枚举、剪枝的复杂方式；处理上直接用数据库系统提供的例程[5][6]而非从数据库交互输入、输出的编程处理。这些一改传统 Apriori 算法将频繁项集一并寻找的运算强度，而是以属性为单位对数据项集进行约简并利用关系数据库中对数据的排序、查找等底层例程提高运算速度。其算法步骤如下：

1） 以项集建立关系表；

2） 对表中每个属性按定义 3 求出其关联度；

3） 按属性的关联度强度分次将相应属性选入测试区；

4）对测试区按定义 1、2 的条件的选择查询操作，得各种属性值组合的支持度、置信度；

5）将满足支持度和可信度阈值的规则输出。

5 挖掘过程

挖掘的数据要有针对性且具有一定的深度和广度。深度是对历史时期的长度,而广度则是分部范围的大小。本例针对三个实体之间近几年多省范围内普通道路的 250 例典型交通事件的分析。限于篇幅且完整地说明问题，对事件的数据类型进行归并和压缩，见表 3。表中记录序号为 N;每个记录相似的事件数量为 n。

表 3 事件综合

Tab.3 Event synthesis

N	时间	地段	典型事件概要	损(死/伤)	n
1	16/07	大连昆明街路	轿车冲撞在人行横道行走中年女子.	1/0	8
2	16/11	苏州月季园东门	程某驾驶车牌撞向路边行人	3/4	6
3	09/09	上海临沂菜场口	苏 K G**5 的轿车经菜场门口失控撞向路边行人.	1/12	7
4	16/11	响水桥路段	鄂 S2***8 轿车程某撞同向推行电动自行车行人.	1/0	16
5	16/05	上海浦东	公交车超越一辆同向行驶的自行车时碾压行人.	1/0	10
6	11/06	湖北武汉	王某驾驶货车刘某驾驶的二轮摩托车发生碰撞，	0/1	7
7	09/01	滑县东留香寨村	魏驾面包车,在撞两行人后,又接连撞 9 人.	8/3	3
8	12/09	南通市郊	郁某驾轿车右侧碰撞同向骑自行车的许某.	1/0	15
9	10/10	南京市郊	秦某驾轿车在找停车泊位时碰到路边的张某.	0/1	12
10	13/11	中山市古神路	吴某驾粤号客车撞前方同向冯某驾驶电动自行车.	1/0	11
11	16/10	丹枫路口	王某驾驶小型汽车光线原因与晨练行人某某相撞.	1/0	18
12	09/09	合肥市郊	严某驾驶该车撞上同向行驶的自行车.	1/0	12
13	16/01	兖州到枣庄	鲁 D***76 的送煤车在左拐弯的时侯碰行人.	0/1	18
14	15/03	深圳机场港平台	杨某驾驶轿车行至机场高架桥离港平台撞向人群.	9/22	4
15	14/2	朝阳、丰台区	薛某和徐某驾驶轿车抢夺单身行人拎包案件 5 起.	0/5	9
16	15/06	淮安盐河路	沈某驾轿车开车门时刮蹭到同向行驶的驾电动车.	1/0	13
17	14/11	西藏街	袁某驾轿车撞倒同向步行的二行人.	2/0	15
18	12/10	田阳县那满镇街	正三轮电动车突然启动冲向行人.	1/0	12
19	14/12	万上线洪田桥	戴某驾赣 JD20**客车光线不强且速度快撞行人，	0/1	14
20	14/12	上万线桐木村	丁某驾赣 C960**轿横突然从停靠车辆前冲出行人，	0/1	7
21	15/1	兵希夏东街	柴某停车打开车门时与驾电动车的王某撞.	0/1	5
22	14/3	溧水开发区	曹某驾助力车从后抢赵某的手袋	0/1	10
23	09/15	张家港百叶路	电动车前轮胎撞到同方行进的行人.	0/1	8
24	16/10	新建丹枫路口	王某驾驶晋号车因光线原因与晨练行人相撞.	1/0	10

设属性分别为事件类型(E)、路级(L)、主体状态(M)、受体状态(O) 、涉事关系(R)和损失结论(C)。

事件类型(E)的属性值分别为:判测不足(a)、失控(b)、分辨障碍(c)、突发事件(d);其中 a、b、c、d 分别涵盖了前述交通事件问题的 1)和 6)、2)和 5)、3)和 4)、7)和 8)项。

路级(L)的属性值分别为: 二级隔离(a) 、一级隔离(b) 、无隔离(c);

主体状态(M): 急左拐(a) 、急右拐(b) 、其它(c);

受体状态(O):面对(a) 、背对(b) 、混合(c);

涉事关系(R): 机.人 (a) 、机.非(b) 、非.人(c) 、 机. 非. 人(d);

损失结论(C)的属性值分别为:严重(a)、重(b) 、中等(c) 、一般(d);

按上述算法第 1)步,根据表 3 的各事件给出的现场记录等资料综合给出各事件关联的关系表，见表 4。设其数据表的名为 Db；

由算法2)、3)对E=a事件集,按定义3，对各属性计算关联度：

rltd(R)=Maxcount(R.x)/Ccount(R)=count(a)/3= 6/3=2 (x=a,b,c);

rltd(C)=Maxcount(C.x)/Ccount(C)=count(b)/4= 4/4=1 (x=a,b,c,d);

rltd(L)=Maxcount(L.x)/Ccount(L)=count(b)/4= 5/4=1.25 (x=a,b,c,d);

rltd(M)=Maxcount(M.x)/Ccount(M)=count(b)/2= 7/2=3.5 (x=b,c);

rltd(O)=Maxcount(O.x)/Ccount(O)=count(b)/2= 7/2=3.5 (x=a,b);

经上各式比较，相关性最强的属性为M和O，将两者选入测试区；

由算法4)按定义1和2，计算属性(M,O)的属性值(b,b)的支持度、可信度：

```
declare @n,@S,@D ;
Sel  @n=count(n) from Db where (E=a);
Sel  @S=count(n)/@n from Db where ((E=a) and (M=b) and (O=b));
Sel  @D=@S*@n/count (n) from Db where ((E=a) and (O=b)) ;
```

@S(支持度)=62/75=0.8;@D(可信度)=(62/75)/(62/75)=1。

按算法 5)输出该支持、可信度。见表 4(右)中式(1)。据此，循环算法 3)~ 5)步，分别得出其它属性值最高支持度、可信度。

类似地，得当E=b、c、d时各事件的属性值典型支持度、可信度。见表4(右)。

表4 事件关系及规则表

Tab.4 Event relation and rule table

N	E	R	C	L	M	O	n	支持度S|可信度D
2	a	a	a	b	b	b	6	(1)(M=b)(O=b):
10	a	a	b	a	b	b	11	S=0.8、D=1;
16	a	a	b	a	c	a	13	(2)(M=b)(O=b)(C=b):
1	a	a	b	b	b	b	8	S=0.6、D=0.7;
17	a	a	b	b	b	b	7	(3)(M=b)(O=b)(R=a)
8	a	a	c	b	b	b	15	S=0.62、D=0.58。
6	a	b	d	c	b	b	7	
23	a	c	d	b	b	b	8	
14	b	a	a	c	b	c	4	(1)(R=a)(O=c):
3	b	a	b	c	c	c	15	S=0.6、D=1;
12	b	a	b	b	c	c	12	(2)(R=a)(O=c)(C=b+)
4	b	b	b	c	b	b	16	S=0.5、D=1。
5	b	d	b	c	c	c	10	
11	c	a	b	b	b	b	18	(1)(M=b)(O=b)(C=b):
24	c	a	b	c	b	b	10	S=0.7、D=1。
19	c	a	d	c	b	b	14	
7	d	a	a	c	c	c	3	(1)(M=b)(O=b):
9	d	a	b	c	b	b	12	S=0.4、D=1;
15	d	a	d	b	b	b	9	(2)(M=b)(O=b)(C=d):
13	d	a	d	c	a	b	18	S=0.5、D=0.8。
20	d	a	d	c	c	c	7	(3)(M=b)(O=b)(R=a):

21	d	b	d	a	c	c	5	S=0.3、D=0.5；
22	d	b	d	b	b	b	10	
18	d	c	b	c	c	c	12	

6 规则分析

对预测不足类事件(E=a)：

表4(E=a)中式(1)为S=0.8、D=1,说明起事体突向右行进且受事体背向起事体的事件占这类事件的80%且可信度达100%，而式(2)、(3)进一步说明出现重损失、机动车涉行人的支持、可信度达(60%、70%)、(62%、58%)；

对失控、驾车报复类事件(E=**b**)：

式(1)说明机动车涉行人群体事件占这类事件的60%且可信度达100%，而式(2)进一步说明出现重损失以上的支持、可信度达50%、100%；

对分辨障碍类事件(E=**c**)：

式(1)说明起事体突向右行进且受事体背向起事体的事件且重损失这类事件的支持、可信度达70%、100% ；

对突发类事件(E=d)：

式(1)说明起事体突向右行进且受事体背向起事体的事件占这类事件的40%且可信度达100%，而式(2)、(3)进一步说明出现一般损失、机动车涉行人的支持、可信度为(50%、80%)、(30%、50%)；

7 分析结论

由上分析可知，对于普通公路各类交通事件，起事体突向右行进且受事体背向起事体所造成重损的事件支持、可信度最高，其中主要是机动车涉人、非机，而非机涉人比例不大。下面结合前述的8个问题给出其主要原因如下：

起事体：为避开与主路物体或突生抛物、行进跟踪精力分散、夜间视觉障碍、靠近抢夺都有急右行进，冲向右侧受事体；报复、失控性起事也多向右侧冲向受事体。

受事体：对后面的情况反应不急且背向受损。

为了避免这两个问题，首先要使受事体有面对事件防范方向和应急处理时间，文献[17]论证了常人对正面和背面的反应时间分别是0.15~0.5S和1~8S，而正向、背向处理紧急情况要有几个数量级的差异；其次若起事体与受事体互相察觉到对方的状态则可规避事件的损失程度。给出以下结论。

1) 行人行进的最佳线路是沿行进方向的道路左侧人行道(或路边)行走。这样就只需观测对面右侧行近的机、非车，规避腹背受敌，因为机、非动车横过道路中线至路左面造成事件的概率小；同时也可有效防范被驾车跟踪、抢夺，因为这类事件要尾随、后位实施，实施者将在实施前就有违规逆行出事的风险，不易操作。

交通规则第六十一条规定行人应当在人行道内行走，没有人行道的靠路边行走。所以这样的安全行进对行人而言只是改变当前靠右侧行走习惯而并不违反交规；

2) 若行人靠左边行走，则机、非动车行进时，面对右侧正向的行人，可对其的结构(尤其夜晚，眼睛反光)、发声、肢体反映、行进意图都会增强分辨度，提高警觉提前防范。

3) 行人以及只靠人力操作的交通工具(自行车、三轮车等)应归并为人力集，该集可按当前行人交规执行，可同样获得上述安全保障。这是因为这些人力交通工具从后面对行人的伤害概率和程度都不高，但机动车从后侧对其所造成的事故远高于正侧，这样同时也可缓解当前的主道压力。

交通规则规定机动车、非机动车实行右侧通行。所以这个实施是需要修改交通规则。

参考文献

[1] 郑长江,沈金星.数据挖掘在高速公路收费数据中的应用[J].中国科技论文在线,2008,3(10):174-176.
ZHENG Changjiang,SHEN Jinxing.Data mining application in highway toll data[J].Science paper Online,
2008,3(10):174-176.(in Chinese)
[2] ARI W,WISNU J,HANIEF AW,et al.Traffic big data prediction and visualization using fast incremental model trees-drift detection(FIMT-DD)[J].Knowledge-Based Systems,2016(93):33-46.
[3] 辛柯俊,吕斌,王忠宇.基于大数据技术的城市交通在线实验环境设计[J].交通信息与安全,2014,32(2):86-89
XIN Kejun,LIANG Biao, GUO Jianhua.Big Data Technology-based Online Urban Transportation System
Laboratory Environment Design[J]. Journal of Transport Information and Safety, 2014,32(2):86-89
[4] 潘晓敏.Apriori 算法实现基于关联规则的交通路段流量挖掘[J].上海工程技术大学学报,2013,27(3):283-288
PAN Xiaomin. Mining Road Traffic Flows Based On Association Rules by Using Apriori Algorithm [J]. JORNAL OF SHA NGHAI UNIVERSITY OF ENGINEERING SCIENCE, 2013,27(3):283-288.(in Chinese)
[5] 贾毅,黄浩丰.SQLServer 在交通事故数据处理中的应用[J].交通信息化,2012,11:131-134
JIA Yil,HUANG Haofeng.Application of SQL Server in Traffic Accident Data Analysis[J]. Traffic Informatization,
2012,11:131-134.(in Chinese)
[6] 张晓博.基于 Python 的 SQL Server 海量数据转移的研究与实现[J],铁路计算机应用,2012,21(2):55-57
ZHANG Xiao bo.Research and implementation of mass data transferred based SQLServer Python[J]. Railway
Computer Application | Railway Comput Appl,2012,21(2):55-57.(in Chinese)
[7] FRE'de'RIC L,PATRICIA D.Speed behaviour as a choice between observing and exceeding the speed limit[J].Transportation Research Part F:Traffic Psychology and Behaviour,2005,8(6):481-492.
[8] NAATANEN R,SUMMALA H.Road user behavior and traffic accident[J].Oxford:North-Holland, 1976,8(5):100-104.
[9] 李山虎.攻击性驾驶行为评价方法研究[D].西安:长安大学,2011.
LI Shanhu.Evaluation method of aggressive driving behavior[D].Xi'an:Chang'an University,2011.
(in Chinese)
[10] NEALE V L,DINGUS T A,KLAUER S G,et al.An overview of the 100-car naturalistic study andfindings[R].
Washington,D.C.,America:NationalHighway Traffic Safety Administration,2005.
[11] 韦华,张伟.中美两国汽车驾驶安全影响因素研究[J].中国安全科学学报,2005,14(9):24-28.
WEI Hua,ZHANG Wei.Study on factors determining driving safety in China and the US[J].China
Safety Science Journal,2005,14(9):24-28. (in Chinese)
[12] 冯忠祥,刘静,李阳阳,等.攻击性驾驶行为选择模型及影响因素敏感度分析[J].中国公路学报,

2012,35(5):210-214.

FENG Zhongxiang,LIU Jing,LI Yangyang,et al.Aggressive driving behavior selection model and its
influencing factors sensitivity analysis[J].ChineseJournal of Highway,2012,35(5):210-214.(in Chinese)

[13] Brehmer B.Variable errors set a limit to adaptation[J].Ergonomics,1990,33(10-11):1231-1239.]

[14]钟足峰,刘伟铭.基于联网收费数据预测交通流量的实现[J].中国管理信息化,2009,12(2):59-60.

ZHONG Zufeng,LIU Weiming.Based on the realizationof the networking toll data to predict traffic
flow[J].China Management Informationization,2009,12(2):59-60.(in Chinese)

[15] 杜勇,李军.交通数据中心数据整合与综合数据库的设计研究[J].交通科计划,2013(2):147-149

DU Yong,LI Jun. Research on the method of data integration and design of integrated database of
transportation data center[J].Transportation Science & Technology, 2013(2):147-149(in Chinese)

[16]黄宇达,王超杰.一种基于改进频繁树模型的智能交通信息关联挖掘方案[J].
计算机与数字工程,2015(2):211-213

HUANG Yuda,WANG Chaojie. An Intelligent Transportation Information Association Mining Scheme Based ImprovedFrequent Tree Model [J]. Computer DigitalEngineering, 2015(2):211-213.(in Chinese)

[17] 赵润栓.平昭等.成年人 WHR 与人体反应速度的相关性分析[J].现代医学,2014,42(2):125-127

ZHAO Runshuan,PING Zhao. correlationanalysis adults'waist hipratio humanreaction speed[J].
Modern Medical Journal, 2014,42(2):125-127.(in Chinese)